Implementing an Open Data Policy

A Primer for Research Funders

Prepared by Greg Tananbaum
On behalf of the Scholarly Publishing and Academic Resources Coalition (SPARC)

Portions of this resource were originally prepared on behalf of the American Heart Association. They are repurposed here with the AHA's permission. SPARC thanks the AHA for its generosity.

What is an open data policy?

Open data policies promote the accessibility and reuse of the raw data generated during the scientific discovery phase. Open data policies typically apply to a range of non-textual materials, including datasets, statistics, transcripts, survey results, and the metadata associated with these objects. The data is, in essence, the factual information that is necessary to replicate and verify research results. "Open data" is the idea that this information should be as freely available as possible, with as few mechanisms of control as possible, to as broad an audience as possible.

Open data policies also usually encompass the notion that machine extraction, manipulation, and meta-analysis of data should be permissible. This extends beyond the human abstraction of facts. Open data polices are not uniform across organizations. However, most research funders will typically require grant applicants to articulate during the application process how they will share their data in accordance with the organization's policies. These statements are called "data sharing plans".

Why should research funders consider adopting an open data policy?

Quite simply, an open data policy aligns with most research funders' strategic goals. These organizations invest in research in order to accelerate the pace of scientific discovery, encourage innovation, enrich education, and to improve the public good. Scientific investigation advances more efficiently and quickly through sharing of methods and results, and the value of an investment in research is only maximized through wide use of this information. At present, the building blocks of science — the data supporting research outputs — are not available to the broadest community of potential users. Internet technologies provide an increasingly cost-effective opportunity to bring these components to a wider audience, and to use these materials in new, innovative ways. An open data policy can facilitate increased discoverability and reusability. This reduces the gaps in the research cycle and makes it easier for interested parties to pursue promising investigative directions. It lessens the likelihood that multiple laboratories will be pursuing duplicative research in siloed environments. It decreases the potential for data miscalculation, misinterpretation, manipulation, and fraud by opening raw results up to the broader community. It also encourages the broadest possible audience to access and build upon research results, a critical advantage for narrow topics that could greatly benefit from the input of interdisciplinary scientists.

From a practical standpoint, an open data policy also demonstrates a tangible return on investment. Many research funders rely on private contributions to support our activities. Disseminating research data in a highly visible manner that promotes sharing, discussion, and follow-up science is a clear way to demonstrate the effective use of donations. It reflects a commitment to good stewardship of the monies with which the funder has been entrusted.

Why should a research funder *require* rather than *request* grant recipients to share their data?

As detailed in a 2013 study published in FASEB, asking researchers to make their data publically available is much less effective than requiring them to do so. To ensure maximum exposure to, and reuse of, data, it is important for research funders to oblige grant recipients to develop a reasonable data sharing plan.

What activities and materials should an open data policy cover?

The exact details of which types of grants and activities should be covered by any open data policy are under the purview of the specific research funder. The guiding principle in making such a determination should be a consideration of what can best accelerate the pace of scientific discovery, encourage innovation, enrich education, and improve the public good. In general, any unprocessed data that is needed for independent verification of research results should be covered by the policy. The data must be accompanied by proper documentation. This documentation, often referred to as metadata, is necessary to allow others to use the data properly and without confusion. Consistent with NIH guidelines, the metadata must provide "information about the methodology and procedures used to collect the data, details about codes, definitions of variables, variable field locations, frequencies, and the like. The precise content of documentation will vary by scientific area, study design, the type of data collected, and characteristics of the dataset."

What constitutes an acceptable data sharing plan?

Given the wide range of projects supported by research funders, no single formula for data formatting, deposit locations, and the like would be universally applicable. However, a broad set of guidelines can be applied to the diverse population of grant recipients. The principles set forth by the Wellcome Trust are helpful in this regard. The Wellcome Trust policy states that data sharing plans should address seven key questions as clearly and concisely as possible:

1. What data outputs will the research generate and what data will have value to other researchers?
2. When will the data be shared?
3. Where will the data be made available?
4. How will other researchers be able to access the data?
5. Are any limits to data sharing required – for example, to either safeguard research participants (e.g., HIPAA) or to gain appropriate intellectual property protection?
6. How will the grant recipient ensure that key datasets are preserved to ensure their long-term value?
7. What resources will the grant recipient require to deliver this plan?

Research funders should require that all submitted data management plans adequately address these seven considerations.

Under what circumstances might a grant recipient opt out of the open data policy?

There may be certain instances in which grant applicants seek to be exempted from the data sharing policy. Broadly speaking, the following categories represent the primary grounds for non-conformance:

- Human Subject Grounds. As the NSF spells out in its exemption guidelines, "[H]uman subject's protection requires removing identifiers, which may be prohibitively expensive or render the data meaningless in research that relies heavily on extensive in-depth interviews." Data sharing cannot violate privacy regulations (e.g., HIPAA) or in any way fail to safeguard the rights of research participants.
- Superseding Regulations Grounds. Governing laws or institutional policies may limit the release of certain data elements.
- Intellectual Property Grounds. Under certain circumstances, data sharing may violate IP rights.
- Financial Grounds. Data sharing should not cause an undue financial burden for the grant recipient.

How can an open data policy be balanced against Health Insurance Portability and Accountability Act (HIPAA) concerns?

As explained by U.S. Department of Health and Human Services guidelines (upon which this section is based), the HIPAA Privacy Rule establishes the conditions under which "protected health information" may be used or disclosed. Protected health information is information, including demographic information, which relates to:

- the individual's past, present, or future physical or mental health or condition,
- the provision of health care to the individual, or
- the past, present, or future payment for the provision of health care to the individual, and that identifies the individual or for which there is a reasonable basis to believe can be used to identify the individual.

Protected health information includes many common identifiers (e.g., name, address, birth date, Social Security Number) when they can be associated with the health information listed above. The relationship with health information is fundamental. Identifying information alone, such as personal names, residential addresses, or phone numbers, would not necessarily be designated as protected health information. Indeed, many of these details are already widely available on the Internet. However, tying this information directly to a health condition or treatment plan, for example, would push this into the realm of protected health information.

The United States Department of Health and Human Services has developed guidance for methods to achieve de-identification in compliance with the HIPAA Privacy Rule. These methods fall into two tracks, (1) a formal determination by a qualified expert; or (2) the removal of specified individual identifiers as well as absence of actual knowledge by the covered entity that the remaining information could be used alone or in combination with other information to identify the individual. It is the second of these paths (deemed the "Safe Harbor" method by HHS) that is typically most relevant to the research funders' open data policy.

The Safe Harbor method of de-identification requires the removal of 18 specific elements from the research data:

1. Names
2. All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: (a) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and (b) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000 3.
3. All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
4. Telephone numbers
5. Fax numbers
6. Electronic mail addresses
7. Social Security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/License numbers
12. Vehicle identifiers and serial numbers
13. Device identifiers and serial numbers
14. Web Universal Resource Locators (URLs)
15. Internet Protocol (IP) address numbers
16. Biometric identifiers, including finger and voice prints
17. Full face photographic images and any comparable images
18. Any other unique identifying number, characteristic, code, except as allowed under the ID specifications (§164.514c)

The researcher responsible for the data should also warrant that he/she does not have actual knowledge that the remaining information could be used alone or in combination with other information to identify an individual who is a subject of the information.

The Department of Health and Human Services provides an FAQ that delves into some of the intricacies associated with the Safe Harbor method.

How quickly should research data be made openly available?

In general, the more quickly data are made openly available, the quicker the impact on further scientific research and discovery. Research funders may decide to allow grant recipients to extract additional value from their data by providing them an exclusive window. In that event, research funders should require that all relevant data to be made publicly available within 12 months of the end of funding period.

What are the implementation costs of an open data policy?

The exact costs will be largely dependent on what grants and materials the policy covers, how the research funder elects to monitor compliance, and a number of other issues. One key consideration is whether the research funder will make additional money available to grant recipients to develop and implement their data sharing plans.

Where should open data be deposited?

Determining the appropriate repositories in which data should be deposited is largely a function of the types of research being conducted and the data outputs generated. It is therefore difficult to identify a list of "pre-approved" repositories that can accommodate all possible datasets. In general, a repository should address the following core considerations in order to be considered appropriate:

- Re-Use. The repository must allow any interested party to freely access the data without restriction on research reuse, ideally via standardized mechanisms such as Creative Commons licenses. This should be codified in the repository's terms of use.
- Security. The repository must articulate how datasets are stored, as well as how any confidential information is protected.
- Stability. The repository must have a clear funding mechanism or business plan to provide reasonable assurances that the data will be available for the indefinite future. It should also have a continuity plan addressing what will happen to the data in the event the repository is discontinued.
- Fee Structure. What is the cost, if any, to deposit data in the repository? Is the payment one-time or recurring? Does the size of the dataset impact the cost? The repository must define its rates and explain how these fees ensure financial stability.
- Subject Focus. There are hundreds of topic-specific repositories in operation at this writing. The grant recipient should endeavor to deposit his/her data in a repository that is appropriate for the subject matter in question. Further, if a repository has emerged within a specific research community as the default resource in that field (e.g., GenBank for DNA sequences), grant recipients should, as a general rule, utilize that repository. This optimizes the ability of others to discover and build upon the data.

- Metadata. The repository must require a depositor to provide sufficient metadata provided to enable the dataset to be used by others. These metadata should be searchable so that repository visitors can easily discovery appropriate datasets.
- File Formats. The repository should be able to accommodate all aspects of the grant recipients' dataset and auxiliary materials, regardless of file type.
- Machine Extraction. The data stored in the repository should ideally be available in a machine-readable and machine-interpretable format.
- Willingness to Accept and Curate Data. Finally, the repository must be willing to accept and curate data submitted by third-party researchers.

What are some examples of acceptable repositories?

The National Center for Biotechnology Information (NCBI), part of the U.S. National Institutes of Health, manages several dozen databases. A complete list may be found here. Deposits to these databases may be made by any researcher free of charge. Internationally, the Wellcome Trust provides a list of repositories that are popular among biomedical researchers.

How does open data relate to open access?

Open access refers to the free, immediate, online availability of peer-reviewed research results, permitting any users to read, download, copy, distribute, print, search or link to the full text of these articles, crawl them for indexing, pass them as data to software or use them for any other lawful purpose . The emphasis is on the research *results*, which typically take the form of scholarly articles. Open data, in contrast, focuses on the factual information that is necessary to replicate and verify research results, including datasets, statistics, transcripts, survey results, and the metadata associated with these objects. Open access and open data are complementary.